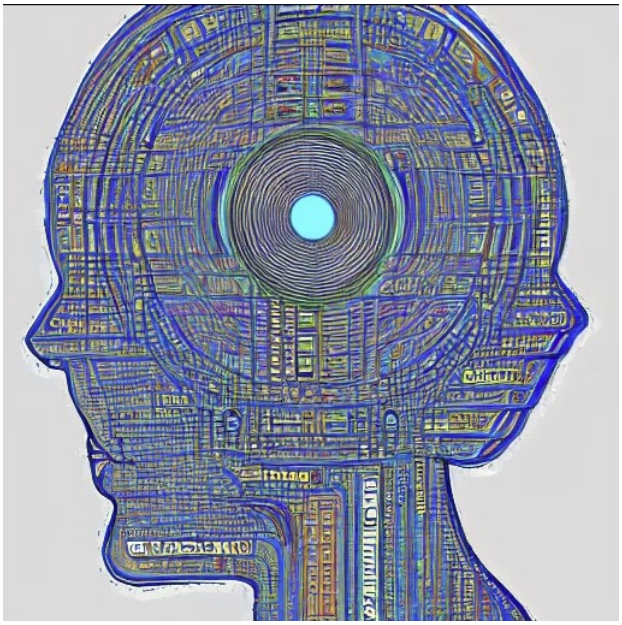

Tools evolving AI systems via experiment management

A survey of machine learning practitioners

Bachelor of Science Thesis in Software Engineering and Management

Carl Vågfelt Nihlmar

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Perception and usage of tools supporting experiment management efforts during machine learning development
-A survey of machine learning practitioners

Supervisor: Samuel Idowu
Examiner: Richard Berntsson Svensson

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Abstract visual of new inteligent system, generated via machine learning

Department of Computer Science and Engineering
University of Gothenburg
Chalmers University of Technology
Gothenburg, Sweden 2023

Tools evolving AI systems via experiment management: a survey of machine learning practitioners

Carl Vågfelt Nihlmar
Department of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gusnihca@student.gu.se

***Abstract*—Artificial intelligence employs machine learning to create intelligent systems. Experiment management tools have been created to support machine learning practitioners in their development efforts relating to the management of artifacts and metadata. Although the technical capabilities of such tools in terms of features have been widely examined, which tools are used as well as the tool´s benefits, limitations and challenges, remain unknown. This paper provides an empirical investigation addressing the questions previously stated. Those interested in gaining a better understanding of the users of these tools, such as tool developers and researchers looking for initial data on this topic, could find the results presented valuable. This was achieved by developing and distributing an online questionnaire to elicit qualitative and quantitative data concerning experiment management tools from 24 machine learning practitioners. Participants reported benefiting from the tools in areas such as reproducibility, time savings, traceability, and result analysis. Reported challenges and limitations of the tools included a lack of features, quality and integration with other systems. Many participants combined tools in order to achieve the desired workflow. The three most commonly used tools were TensorBoard, MLFlow, and SageMaker. The empirical contributions of the survey improved the understanding of experiment management tools from the perspective of machine learning practitioners. The data can be leveraged towards building better supporting tools for AI development and serve as a basis for further research in related areas.**

***Index Terms*—Experiment management tools, machine learning, experiment management, experiment tracking, artificial intelligence.**

I. Introduction

Machine learning (ML) is considered a subset of the artificial intelligence domain as its purpose is to develop intelligence via different types of learning [1]. ML is currently delivering significant value within a wide variety of applications due to the remarkable advancements in the field. When disruptive technologies emerge, such as the internet, cloud computing, or ML, organizations must address the inherent challenges faced when attempting to take advantage of such technologies.

One inherent challenge of ML, on a management level, is its need for extensive skills in software development, mathematics, statistics, and data. However, such expertise is often lacking in conventional software teams. The requirement for diverse skill sets has made it increasingly challenging to achieve effective team collaboration and coordination. [2]. This difficulty has been attributed to the difference in team members' training background, e.g., an engineer's training vs. that of a data scientist or a mathematician [3]. Several technically challenging aspects of ML exist, three of which are exemplified below.

First, intelligent components built using ML are often entangled in complex ways. This entanglement leads to non-monotonic error propagation. Therefore, changes that lead to improved performance in one component could cause multiple other components to have decreased performance. The CACE principle, "changing anything changes everything", stems from this entanglement trait of ML [4]. Second, Amershi et al. [5] state that handling data in terms of discovering, managing, and versioning within the context of usage in ML applications is much more complex than other types of software engineering. Third, according to D. Sculley et al. [4], [6], ML systems are more prone to accumulate technical debt than non-intelligent systems due to distinct risk factors such as data dependencies, hidden feedback loops, and multiple system-level anti-patterns.

An experiment management tool (EMT) is designed to assist ML practitioners with their experiment management efforts. It achieves this by allowing its users to make queries regarding previous ML experiments. For example, a user's query could search for experiment results or specific components of runs to gain insight into how the results were achieved. EMTs offer several benefits by addressing previously discussed challenges of ML. Examples of such benefits include: (i) from an organizational perspective, team collaboration can benefit by knowing who was involved in running which experiments, (ii) identification of questions for consideration of efficient technical debt payoff as studied by D. Scully et al. [6], three of which are listed below:

1) How precisely can the impact of a new change to the system be measured?
2) Does improving one model or signal degrade others?
3) How quickly can new team members be brought up to speed?

An EMT that can query and visualize information from previous and current runs could be significantly helpful in the aspects touched by all three questions above, and (iii) the same abilities of an EMT could be used to mitigate the complexity of data management and the entanglement described by the CACE principle. Additionally, experiment management and EMTs have been shown to reduce redundant development efforts and allow more efficient testing and debugging [7]. Such positive effects can be extrapolated to substantially impact a project's progress and overall costs in time, human resources, or money.

A. Problem domain and motivation

The need for asset management and EMTs have been well documented in previous literature [8]. ML practitioners have been interviewed in previous studies with fundamental questions like "Can't we go back to see how we used to be doing?" referring to aspects of a previous experiment [7]. ML practitioners described their solutions as ad hoc or myopic and reported them to include emails, notes, file-naming conventions, databases, VCS, and Git. The same practitioners state that a critical negative aspect of these solutions is the ratio of their value to the time and effort required to implement them. As a result, they would often have to compensate for the shortcomings of the solutions with face-to-face meetings.

As identified by Berger et al. [9], something lacking from academia is the empirical data on ML practitioners' usage of EMTs in terms of which tools they use and their perspectives and opinions of these tools. This lack of knowledge hampers the ability of EMT developers to improve their tools relative to their user base needs. Additionally, it impairs new grounds for further research and development of industry practices relating to experiment management and EMTs.

B. Research goal and research questions

The survey conducted in this study is focused exclusively on EMTs and aims to fill the established research gap by eliciting information via an online questionnaire. The questions asked in the survey via the questionnaire were based on the following three research questions:

- RQ1: What experiment management tools are used?
- RQ2: What are the benefits of using the tools?
- RQ3: What are the major challenges and limitations of using the tools?

Limitations focus on the lack of features and technical capabilities that obstruct a tool's utility. Similarly, challenges are oriented toward aspects that make the tool difficult to use. The results presented in this paper could be found valuable by researchers needing initial data on this aspect of experiment management and EMT developers wanting to understand their user base better.

C. Structure of the Article

The next sections of the paper are organized as follows. First, related work and background are discussed in Section II. Next, Section III describes the methodology, the questionnaire´s design, distribution, and the analysis used. The results of the analysis are reported in Section IV. These results are then discussed in Section V, with threats to the results' validity and opportunities for future research. General conclusions are finally drawn in Section VI.

II. Background and Related Work

This section discusses terms and concepts that will allow for a deeper understanding of both EMTs and the content primarily presented in Sections III, IV, and V. It comprises four subsections where the first three are used to build a frame for the fourth: (i) ML workflows, (ii) ML assets, (iii) ML concerns, and (iv) foundations of EMTs.

A. ML workflows

The development process of ML is described as exploratory as it iteratively utilizes multiple series of experiments. This development style is closer to that used in data science than traditional engineering [10]. A conducted experiment is often referred to as an experiment run. After a run, the experiment and its results are carefully reflected upon for insights into what modifications can evolve the system. The process then repeats with subsequent experiment runs. The runs often reach large quantities due to numerous iteration cycles [8]. Extensively repeated experiment runs are not exclusive to the development phase but also extend to the maintenance phase, where re-training via more runs is used to keep the system performing as desired [11].

A few ML methodologies have been more prominently adopted than others by academia and industry. In particular, two methodologies exemplifying this are the Cross-Industry Standard Process for Data Mining (CRISP-DM) [12] and the more recent Team Data Science Process (TDSP) [13]. A case study focusing on the ML workflows at Microsoft was conducted by Amershi et al. [5]. The researchers found that Microsoft used a workflow with commonalities to CRISP-DM and TDSP. Using a case study approach, they presented nine consecutive stages to represent all phases encapsulating the life cycle of an ML project. The first of these nine stages was the "Model requirements" stage, followed by three consecutive data-related steps: (i) data collection, (ii) data cleaning, and (iii) data labeling. The remaining five steps included feedback loops which could lead back to previous steps. These remaining five consecutive steps were more strongly related to models rather than data and comprised of: (i) feature engineering, (ii) model training, which could lead back to the feature engineering stage, (iii) model evaluation, which could lead back to any of the previous stages relating to data or models, (iv) model deployment, and (v) model monitoring, which could lead back to any of the previous stages relating to data or models.

B. ML assets

From every experiment run, various assets can be derived. Examples of these assets are artifacts relating to data, models,

software, dependencies, and metadata of the experiment and its results [14]. The discipline of storing these assets so they can be queried at a future point is called experiment management [9]. This type of management enables leveraging information from previous runs to, for instance, better evolve a system's performance or to efficiently onboard a new team member. Growth in an ML project in terms of, for example, scope, models, data, the number of runs, or collaborators tends to compound ML challenges. However, this can be offset by experiment management; consequently, the utility potential of experiment management increases with the growth of an ML project. The assets that can be produced or derived from an ML run are many and often difficult to classify. Idowu et al. [9] provides a comprehensive taxonomy that defines experiment assets more clearly. The taxonomy establishes four different categories to enable the classification of assets: (i) support software assets (including source code, notebooks, and parameters), (ii) resources (including datasets, models, and generic resources), (iii) various metadata (including experiments, code, data, and models), and (iv) ExecutionData (including dependencies, jobs, ExecutionMetadata).

C. ML concerns

During an ML project, modifications made between runs essentially create a new version of the experiment, similar to how in traditional software engineering development, new versions of a codebase are created as developers modify it. Versions of a codebase can be managed using mature version control tools such as Git. However, two critically differentiative factors between the development of ML and traditional software engineering, as it pertains to leveraging knowledge from previous versions of a project, are listed below. Firstly, the number of versions produced is usually significantly higher in an ML project due to more frequent iteration cycles, i.e., the experiment runs. Secondly, the complexity of each version is higher in an ML project due to the number of different assets that can be modified in each iteration being significantly higher in ML development. Consequently, these two factors have caused some aspects to be more difficult to manage. Exemplifying such aspects are the following prominent ML concerns listed below, along with a question to understand the scope of each concern better:

1) Auditability [15], [16]: How well can audits, internal or external from third parties, be made regarding the results and how results were achieved?
2) Collaboration [2], [3]: How well can ML practitioners collaborate in an ML project?
3) Interpretability [17], [18]: How well can the underlying logical principles of a model be understood?
4) Reproducibility [19]–[21]: As per Tatman et al. [22], how well can the results of an experiment be recreated in a new experiment using the same input data, models, and analysis?
5) Replicability [23]: As per Tatman et al. [22], how well can the results of an experiment be recreated in a new experiment using the same models and analysis but with new input data?
6) Traceability [24], [25]: To what degree are aspects relating to data and models documented?

D. Foundations of EMTs

The inadequacy of traditional tools and ad hoc solutions to not adequately address the previously discussed challenges and concerns of ML lays the foundation from which the EMT class of tools has emerged. Figure 1 illustrates how EMTs can achieve this new type of support by integrating the ML projects running the experiments. This support is, as illustrated, enabled by allowing the tool to extract and store relevant assets to allow users to retrieve and visualize them.

E. Related work

Multiple surveys review EMTs [26]–[29]. Literature comparing and evaluating EMTs and related ML support systems has also been published [9], [19], [30]–[32]. Schlegel et al. [31] conducted a comprehensive survey on the functional scope of over 60 systems and platforms. They discuss ML support systems, including EMTs, as an essential building block to managing ML concerns. By reviewing the literature, they derive functional and non-functional criteria that they then use to assess the scope of their selected systems and platforms. Weißgerber et al. [30] delivered an open science-centered process model for machine learning research based on the author's review of features of over 40 tools, platforms, and standards. Based on these findings, they list the tools found to be central to the paper's research process. Isdahl et al. [19] surveyed several systems' abilities to enable the reproducibility of empirical ML results. They propose a quantitative method that can be used to assess the system's reproducibility support and then apply this method to evaluate the state of reproducibility supporting tools. All four studies [19], [28], [30], [31] involved EMTs, but tools that fall outside of experiment management solutions, such as model and pipeline management, are also included. Idowu et al. [9] surveyed the capabilities of 17 tools within the experiment tracking and management domain. Additionally, they provided a feature model that describes EMTs' variability and commonalities, which can be used to assess current and future tools. Similarly, Quaranta1 et al. [32] evaluated 19 tools to create a taxonomy of what support tools can provide concerning ML reproducibility.

By contrast, the research questions presented in Section I, are only concerned with the tool's features or capabilities through the lens of their users, which separates the study from surveys with purely a technical perspective. Hill et al. [7] conducted interviews at a company to better understand ML practitioners in the context of what skills they need, what tools they use, and what problems they face. Data on ad hoc experiment management solutions, skills needed for ML development, and ways of working are elicited from interviewees and presented in the study. However, it lacks the inclusion of the EMTs as this class of tools has evolved significantly since the study was published in 2016. Zhang et al. [2] conducted an online

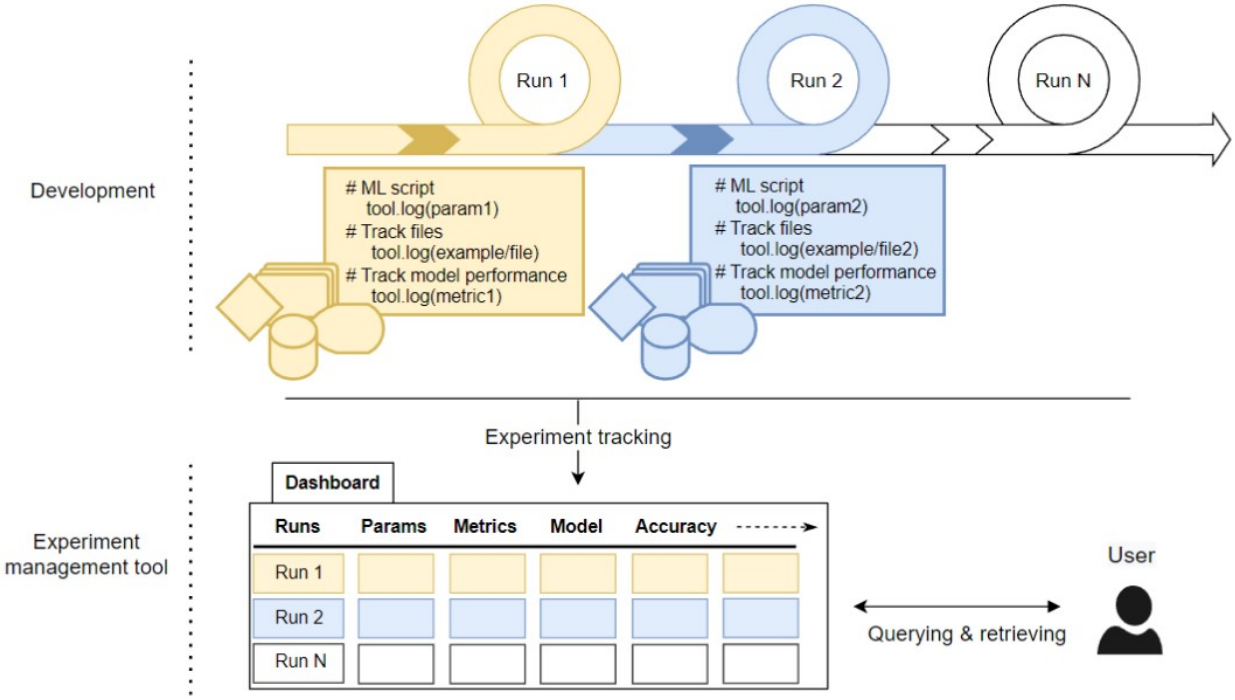

Fig. 1. Illustration of how an EMT can support its users via interacting with a machine learning project.

survey in which they elicited information from 183 participants working in data science concerning what tools they use, what roles they have, and their workflows, all from the perspective of collaboration. The study briefly mentions Google Colab, a tool with EMT features, and how it is used for collaboration. Despite the literature presented in this section, the research questions, as defined in Section I, have not yet been sufficiently addressed using empirical data.

III. Research Methodology

An online questionnaire survey was deemed most appropriate to elicit information from ML practitioners regarding EMTs. The target population was ML practitioners, i.e., researchers performing ML experiments. The insights from multiple discussions with members of the international research group named EASElab influenced the survey's scope and the questionnaire design. Informal interviews with ML practitioners regarding experiment management were conducted at Gothenburg Artificial Intelligence Alliance (GAIA). The knowledge gained from these interviews was then leveraged towards the survey and questionnaire.

The research scope of the survey was iteratively narrowed until it was finalized into three research questions defined in Section I. As recruiting survey participants is a lengthy process, the questionnaire was distributed as soon as the scope was deemed sufficiently defined. The guidelines for a questionnaire-based survey presented by Linaker et al. [33] were used as a reference guide throughout the survey.

A. Questionnaire design

The questionnaire was designed to take ten minutes of the participant's time, with minimal mental resources to complete while still eliciting sufficient information. All participants had to respond via checkboxes for consent and voluntary participation in the survey. If consent was not given, the participant could not continue answering the survey questions.

The questionnaire constituted of six sections: (i) participants performing ML experiments, (ii) participants who do not use EMTs, (iii) participants who use EMTs, (iv) limitations and challenges of EMTs, (v) participant info, and (vi) contact information. The first was a control question to see if the respondent was qualified enough to answer the questionnaire by asking about the person's history with ML experiments. Through these six sections, there were two separate tracks a participant could take: (i) path one for those who reported they did use EMTs, and (ii) path two for those that did not use EMTs. Participants that reported not using EMTs were not asked the questions in sections named "Participants who use EMTs" and "Limitations and Challenges of EMTs" but were instead asked the questions in "Participants who do not use EMTs". This structure made it possible to tailor the questionnaire to suit all types of participants relevant to the survey. The survey instrument elicited qualitative and quantitative data from participants. Both open, closed, and partially-structured questions were used. The questionnaire was peer-reviewed multiple times throughout its creation by research group members before distribution.

In order to validate and assess the survey instrument, dry runs were conducted with the recruitment of potential pilot

participants undertaken via convenience sampling. The dry runs commenced when the peer-reviews deemed the survey instrument to be matured. The participants were then sent an online questionnaire asking them to: (i) report any issues and suggest potential improvements, (ii) experiment with tracks for both 'EMT users' and 'non-EMT users', and (iii) measure and report the time taken to respond to each path. The response generated to the three points, in addition to the level of familiarity of participants with ML and EMTs, was then used to conduct unstructured interviews with each participant to elicit information useful for further enhancing the questionnaire.

Some key changes made to the survey instrument based on the interviews include: (i) formulating the cover letter's value proposition to promote a higher participation rate, (ii) enhanced understanding of the survey instrument via improved clarity of questions and section descriptions resulting in the reduced completion time of the questionnaire while eliciting information with increased accuracy, and (iii) adapting answer options, e.g., changing an open-ended textbox-type question to a structured multiple-choice question to elicit more specific and accurate information. A participation time of ten minutes was observed during the pilot study, which validates the general design of survey instruments in this aspect. In total, five people participated in the dry runs, which ran over a period of one week. The survey instrument was built in google forms, and the participants' responses to the questionnaire can be found here: https://cutt.ly/P2dky7c.

B. Sampling strategy and questionnaire distribution

The non-probabilistic convenience sampling method was utilized to find the participants for the survey by following the approaches as outlined in this section. As this study aims to gain insight into ML practitioners' usage of tools and opinions, the convenience sampling method was optimal due to its inherent advantages, including reduced time requirement, high speed, and low operational cost.

For the distribution of the survey instrument, contact information was collected from ML practitioners attending GAIA via personal interactions with those willing to participate or who would be able to distribute the questionnaire to their friends and colleagues. Similarly, all research group members distributed the questionnaire via email to colleagues. Additionally, posts were made on six different ML-related forums and five ML-focused LinkedIn community groups. The forums used for such posts were focused on ML and, therefore were expected to have community members interested in the topic and, by extension, an interest in experiment management tools. The names of the forums and the newsletters are discussed in the subsequent subsection III-C and can be found via the following link: https://cutt.ly/o3kqEa5.

The cover letter included in the questionnaire encouraged the participants to refer the survey instrument to other potential participants. This type of referencing was used to enable the effect of snowball sampling [34] and further increase survey participation. Additionally, the cover letter was designed to promote participation by providing an option to receive a summary of the questionnaire's results. The survey was open for a period of six weeks to collect data, with a total engagement of 24 participants undertaking the survey.

C. Distribution issues

A high participation rate was anticipated to be difficult. However, practically it was observed to be significantly challenging. The incentive of receiving a result summary report and the opportunity to contribute to improving the experiment management field seemed insufficient. An example to illustrate this issue is a forum post made on Reddit (see link in Section III-B for details), which gathered over 2500 views, with only 2 of them undertaking the survey. Additionally, fifteen different ML-related newsletters were contacted, with no positive outcome for distributing the survey further. It was observed that personal connections had a higher success rate in attracting participants than forum posts. The tactic of sending reminder emails to all who had received the initial invite to participate was used with some success as some participants reported having forgotten to participate after the initial invitation.

D. Data analysis

The quantitative data were analyzed using descriptive statistics, and the qualitative data were analyzed using thematic analysis with deductive coding following the guidelines provided by Braun and Clarke [35]. The sample size is relatively small, as only 24 individuals participated. The implications of the sample size are further discussed in Section V.

IV. Results

This section presents the results from the questionnaire instrument as described in Section III. The first subsection includes the findings from questions that all participants answered. The subsequent two subsections present results from questions asked to those who did not use EMTs, followed by results from those who did use EMTs. Some of the questions allowed participants to make multiple choices. The term "relative selection" is used in illustrations for multiple-choice questions. The calculation of the relative selection was based on each answer´s frequency count divided by the total number of participants responding to the question. Some of the questions elicit information using Likert scales where participants can agree or disagree with a written statement presented to them. These questions focus on participants' sentiments and include a linear trendline for a more straightforward overview. Open-ended questions were analyzed using thematic analysis as described in Section III. Figure 13 in the appendix is a visual representation of the themes and codes per each open-ended question.

A. Survey participants

Table I lists a grouped representation of the participant's years of experience working with ML experiments. The participant with the most experience was 20 years, while the participant with the least experience was 1 year. Moreover,

the average participant's years of experience was 5.8 years.

TABLE I
GROUPED YEARS OF EXPERIENCE WITH ML EXPERIMENTS.

Years of experience	Frequency	Percentage
1–3	9	45
4–7	7	35
8–11	2	10
> 11	2	10

Together the respondents represented 14 different industry domains, see Figure 2 with technology being the most common, followed by health, finance, and education.

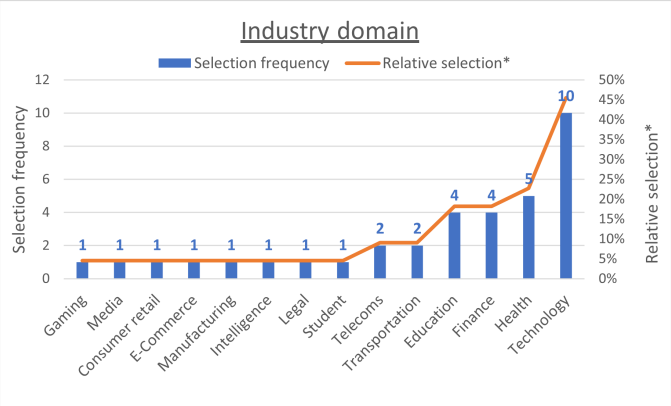

Fig. 2. Illustrates responses to the question "In which domains do you currently work?".

Figure 3 illustrates the roles which participants found most accurately describe their current work. Here "data scientists" was the most commonly reported role with 61% of the participants selecting it, followed by "ML engineer" which was selected by 52% of the participants. Further, when participants were asked if they used EMTs, 30.4% responded that they did not use such tools, and the remaining 69.6% responded that they did use such tools.

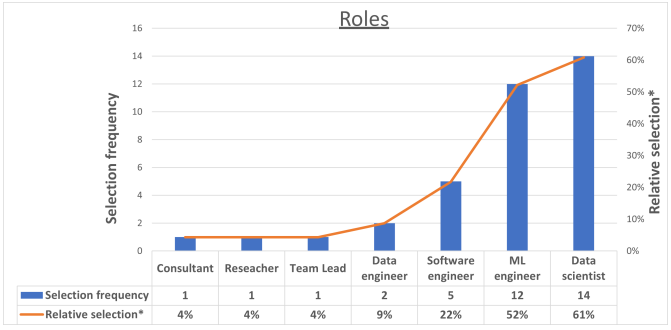

Fig. 3. Illustrates responses to the question "What are your current roles?".

B. Participants not using EMTs

This subsection presents the results from the follow-up questions asked exclusively to the 30.4% of the total participants who previously stated that they did not use EMTs. When asked if they were aware of the existence of EMTs 57.1% of the respondents answered "yes", and the remaining 42.9% responded "no".

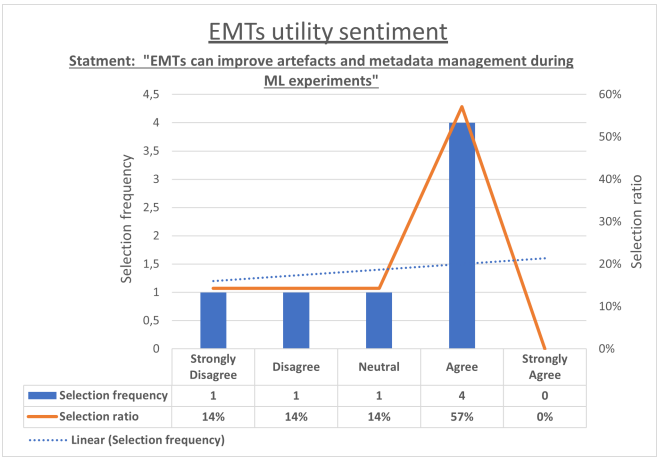

Fig. 4. Sentiment towards EMTs potential utility.

Participants were then questioned on how they perceived EMTs' utility as exhibited in Figure 4. They were asked to respond to the following statement, "EMTs can improve artifacts and metadata management", via a Likert scale ranging from strong disagreement to strong agreement. Despite not using the tools, 57% of participants agreed with the statement, and the overall sentiment showed a positive attitude toward the tool's potential utility.

When participants, who previously stated they were aware of EMTs' existence, were asked why they did not use such tools via an open-ended question, three unique themes emerged: (i) use-case (no need for dedicated EMT), (ii) migration (switch from current tools too cumbersome), and (iii) tool attributes (learning curve). Participants were then asked what solution they implemented for the management of ML assets, if any. Figure 5 presents the results for this question, which show naming conventions to be most utilized. Only one respondent reported that they do not manage versions.

C. Participants using EMTs

This subsection presents the results from the follow-up questions asked to the 69.6% of the total participants who previously stated that they do use EMTs. From a list of 37 different EMTs, the participants were asked to select which they used, if any. The participant's tool selections are illustrated in Figure 6. TensorBoard was the most used tool, with seven participants reporting using it.

As it is possible to use more than one tool, participants were asked to state how many tools they use. The results of this question are presented in Figure 7, which shows that using a single tool as well as using three tools were the most frequently reported options.

Participants were asked to agree or disagree with four statements regarding the utility of EMTs. The results show that those who use EMTs are very positive towards the tool's usefulness, stating that they provide benefits in multiple aspects.

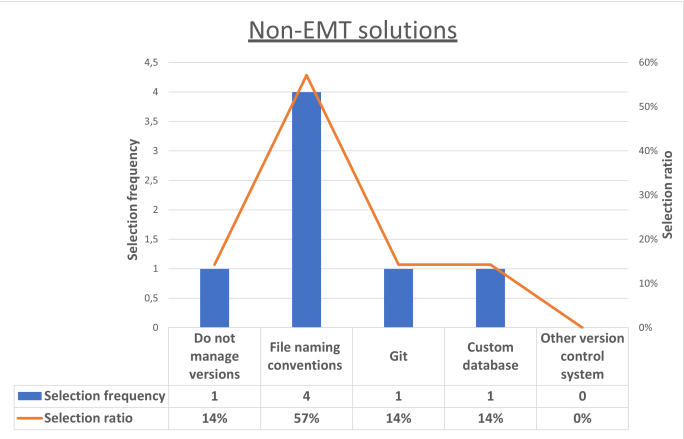

Fig. 5. Illustrates responses to the question "How do you manage versions of your experiment artefacts and metadata?".

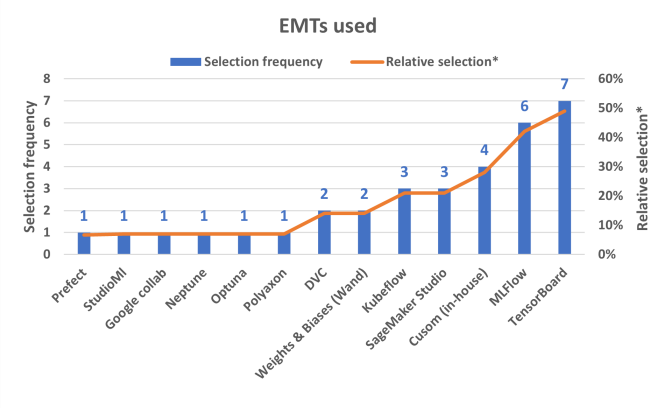

Fig. 6. Illustrates responses to the question "If yes, which of the following experiment management tools do you use?".

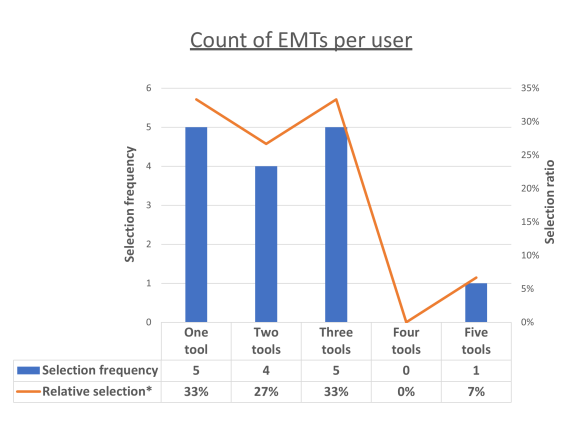

Fig. 7. Tools per user.

This positive attitude was most prominent in the responses to whether it makes the practitioners perform experiments more efficiently and if they provide a benefit over the alternative of not using an EMT, as shown in Figure 8.

Participants were then asked to clarify in which aspects if in any, they experienced these benefits. Reproducibility was the aspect in which most participants reported benefits, closely followed by time-savings, as shown in Figure 9.

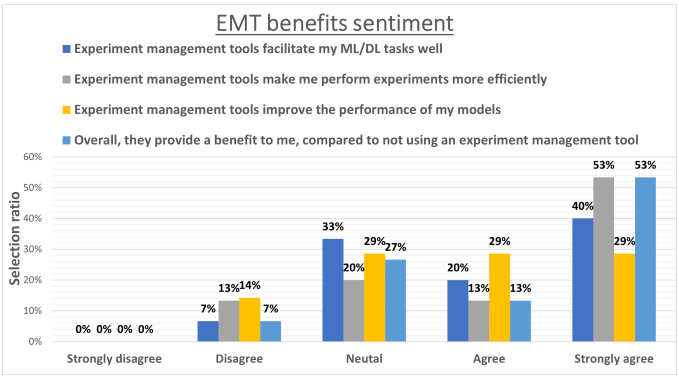

Fig. 8. Sentiment read of EMT benefits.

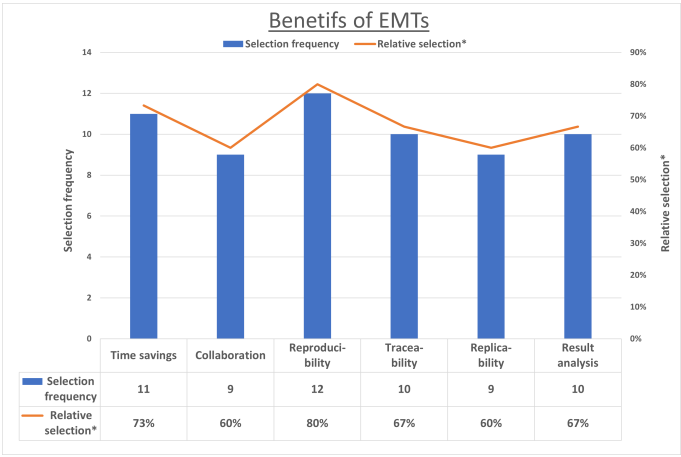

Fig. 9. Illustrates responses to the question "If applicable, where do you see the benefits/values of the EMTs that you use?".

Another question using a Likert scale was then used to gauge to what degree EMT users felt that the limitations of their tools were affecting their experiments. The scale once again ranged from strong disagreement to strong agreement. The sentiment was favorable towards EMTs as respondents disagreed with the statement, as shown in Figure 10.

Further, an open-ended question was posed to learn more about the limitations the participants had experienced. From the responses to the open-ended question via thematic analysis, three main themes could be derived: (i) tool attributes (cost), (ii) features (lack of features, lack of feature quality, user experience), and (iii) external tool integration (visualization tools, custom pipelines, version control, databases).

When asked about the challenges participants had experienced concerning EMTs, using another open-ended question, three similar themes emerged: (i) tool attributes (learning curve), (ii) features (lack of features, lack of feature quality, tool documentation, user experience), and (iii) external tool integration (visualization tools, custom pipelines, version con-

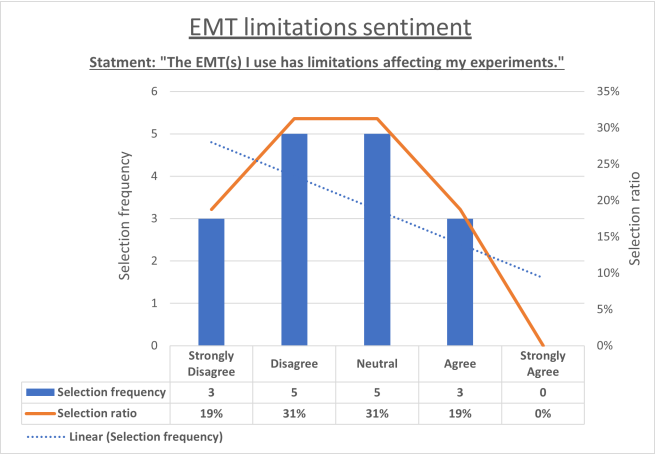

Fig. 10. Sentiment read of EMT limitations.

trol, databases).

V. Discussion

This section intends to answer and discuss results through three subsections, each relating to one of the research questions introduced in Section I followed by a discussion on threats to validity.

A. Answer to RQ1: What experiment management tools are used?

The EMTs that participants used were presented in Figure 6. The five most popular of these were: TensorBoard, MLFlow, KubeFlow, SageMaker, and in-house custom-built tools. EMTs are often used in combination with each other, creating a toolchain when practitioners employ a combination of tools rather than a single tool. The average participant used about two tools per user. Additional information could be derived from the raw data generated from the responses to the multiple-choice question of which tools participants use. Estimating how often each tool was used with another tool provides insight into how frequently it is combined in the workflow with other tools. Figure 11 represent the results of such calculations and includes all tools reported as being used in combination with other tools and thereby had a pairing frequency count higher than one. Higher pairing counts indicate a tool was included in, for example, shorter toolchains with many participants or alternatively in longer toolchains with lesser participants, i.e., toolchains with many tools. Both the alternatives could be attributed to a tool's high pairing count.

Furthermore, the pairing count can be used for additional insights into what tools are most commonly combined with others on a relative basis. Dividing each tool's pairing count, as illustrated in Figure 11, with the number of participants reported using it, as illustrated in Figure 6 yields the result presented in Figure 12. Higher numbers indicate a tool was more frequently used by participants with longer toolchains.

The relative pairing counts are noteworthy as DVC, a tool with a comparatively narrow feature scope, is followed by SageMaker and kubeFLow, two tools focusing on the full ML life cycle. These results indicate that ML practitioners' tool selection is complex and varied. The findings also open up broad ranges of avenues for future work, which are discussed in Section VI. The tool found to be combined with other EMTs most often relative to how many participants reported using it was found to be DVC.

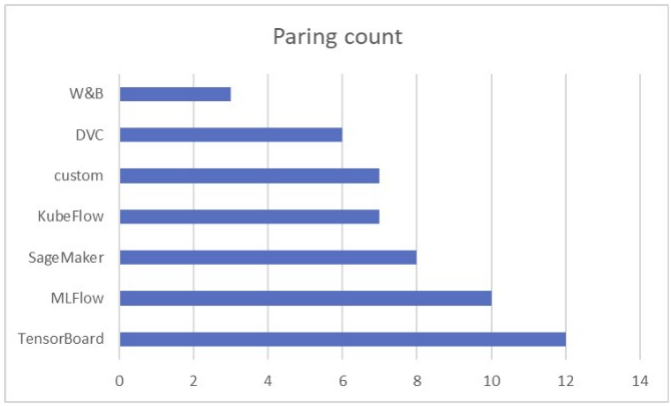

Fig. 11. Paring count of all tools combined more than once with others illustrating how often each tool was combined with other tools.

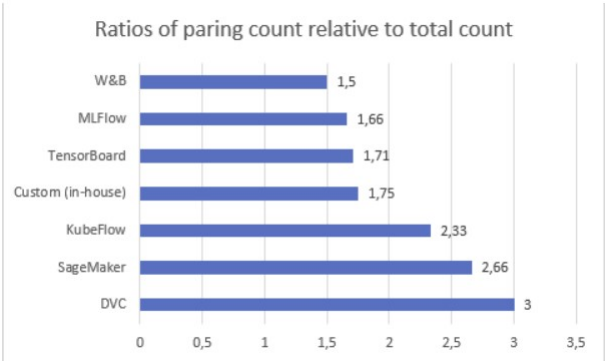

Fig. 12. Relative Paring count of all tools combined more than once with others illustrating how common each tool was to combine with others relative to the number of participants reported using a specific tool.

B. Answer to RQ2: What are the benefits of using the tools?

Both categories of participants who do and do not use EMTs demonstrate an overall positive sentiment toward the utility an EMT tool can provide, as illustrated by the data in Figures 4 and 8. In Figure 9, benefits were reported across six particular aspects relating to ML. The least frequently reported categories of benefits were collaboration and replicability, although a significant share of 60% out of all EMT-using participants stated that they did experience their EMT as providing value in these two areas. Reproducibility was found to be the aspect in which most participants experienced benefits. Being able to reproduce results and experiments others have created is an integral part of conducting research within academia. Due to the challenges of ML as described in Section I and Section II, research publications involving ML have had difficulty

achieving reproducible experiments and results. The issue of reproducibility is so pronounced that Gundersen [36] referred to it as a crisis. It is reasonable to conclude that EMTs could play a significant role in addressing the reproducibility issue as 80% of EMT users reported that their tool benefits them in this aspect. Other notable findings more closely related to concerns valuable to industry rather than academia are time savings, collaboration, and result analysis. All three aspects could have an essential impact on industry ML projects' success, and all three aspects saw at least 60% or more respondents reporting experiencing benefits in these areas.

C. Answer to RQ3: What are the tool's challenges and limitations?

The thematic analysis of the data concerning challenges and limitations share the same common themes and differ only slightly in the codes that make up the themes. For example, the learning curve and tool documentation were two aspects uniquely reported as challenges that made the tools difficult to use. In terms of limitations, the cost was the only aspect that could not be found reported as a challenge. The term quality, used in the thematic analysis, describes bugs, robustness, and reliability issues. At the same time, the lack of features is simply the absence of capability in some aspects. The lack of quality and capability in terms of tool features indicates that EMTs as a class of tools are relatively new and are yet to reach maturity, as also indicated by the findings in RQ1 where most participants use more than one tool to satisfy use-case needs. Furthermore, integration is widely reported as both a challenge and a limitation, and should be considered a key aspect of an EMT. ML practitioners have strong interest in being able to integrate their EMTs with a broad range of independently intricate systems, such as databases and visualization tools. Therefore, providing high-quality and capable integration options with EMTs though particularly challenging for tool developers, would be greatly appreciated by users if delivered.

D. Threats to validity

This section discusses the threats to validity and reliability in relation to the current survey.

1) Construct validity threats: To ensure that the questions of the survey instrument would elicit data that could be used to answer the research questions as defined in Section I peer-reviews were used. Additionally, the research questions were designed and reviewed to ensure they were clear and focused enough not to threaten the construct validity of the survey. To avoid participants misinterpreting any part of the survey clarifying explanations and examples were provided in instances where this was deemed a risk. Efforts to mitigate the risk of survey participants misconstruing any part of the questionnaire were made during each follow-up interview conducted after every dry run.

2) Internal validity threats: The first research question of this survey, regarding which tools are used, is less prone to outside influences than the other two research questions as they elicit sentiment-related data. Additionally, the fact that only one research method, an online survey questionnaire, was used adds to this threat. Sentiment data was elicited more than once using open and partially structured questions to mitigate this threat. The decision to use online forums made the survey more prone to receive malicious or fake responses, which could harm the internal validity. This threat was mitigated by having all responses analyzed manually to detect such responses. No responses were deemed to fall into such categories. Furthermore, to strengthen internal validity, each participant was asked whether they had performed ML experiments as the first question of the questionnaire. If the participant provided a negative response to the question, they were unable to continue with the survey. This was used to ensure that participants had domain knowledge.

3) External validity threats: The limited sample size is a threat to validity that must be considered. The challenges of increasing the sample size and all the steps taken to mitigate the issue are more extensively described in Section II.

According to Searle [37], convenience sampling is prone to the issue of having too many survey participants similar to our own social and cultural groups. By extension, the attempt to achieve the snowball sampling effect via the cover letter, the usage of colleagues and personal networks, and the GAIA conference attendance to find survey participants all add to the threat of external validity.

Babbie [38] explains how the inability to control a study's sample distribution in terms of representativeness is at the core of the issue when utilizing convenience sampling. In an attempt to detect the over-representation of a particular type of participant in the survey, the data elicited on roles, industries, and years of experience can be assessed. When considering the distributions in these three areas, it becomes clear that some segments in some categories hold a significant majority, technology in industries and data scientists in roles. However, considering that these two questions were multiple-choice questions combined with the fact that participants are ML practitioners, such over-representation in these categories could exist even in a healthy distribution of survey participants and is, therefore, not necessarily a cause for concern.

4) Reliability threats: During data analysis, no inconsistencies were found in the responses, including questions regarding sentiments toward EMTs. The consistency in similar-themed questions producing similar results strengthens reliability. A link to the questionnaire used by survey participants is provided in this report to promote reliability through replicability and reproducibility. Limited sample size also meant less data requiring thematic analysis, which lowered the amount of data prone to the threat of author bias. To further mitigate this threat, fellow researchers' opinions of the data were considered during thematic analysis.

VI. Conclusions & Future Work

This paper presents a survey based on an online questionnaire that elicited information on ML practitioners' use and opinions of EMTs. By analyzing the qualitative and

quantitative data, insights into which EMTs are used, their benefits, and their deficiencies in terms of challenges and limitations, could be derived from ML practitioners. The sentiment towards the tools was overall positive. A majority of the participants reported benefiting from the tools in areas such as reproducibility, time savings, traceability, and result analysis. Reported challenges and limitations of the tools included a lack of capability and quality in terms of features and integrations with other EMTs and non-EMT systems. Many participants combined tools in order to achieve the workflow they desired. The average participant using EMTs combined the use of two such tools and the three most commonly used tools were TensorBoard, MLFlow, and SageMaker. The tool found to be combined with other EMTs most often relative to how many participants reported using it was found to be DVC. I believe that the findings presented in this paper will promote increased interest from academia and industry and consequently improve this class of tools and the development of future ML systems.

Concerning future work, numerous approaches could be pursued to learn more about EMTs. Replicating this survey with a larger sample size could be used to determine the validity of the results and conclusions drawn from the data presented in this survey. Valuable insights, especially as it pertains to expanding on the findings regarding RQ1, could come from eliciting data on the context from which decisions and opinions around EMTs are made and formed using additional data sources such as interviews. A deeper understanding of the context, e.g., team size, use-case, could enable more fine-grained comparisons of which EMTs are used when and why. That data could enable many avenues of investigation such as sentiments towards the tools utility and their challenges by applying the research questions RQ2 and RQ3 as per I but within the new contexts identified.

VII. Acknowledgments

This study would not have been possible without the support and guidance of my supervisor Samuel Idowu. For your efforts throughout this thesis, you have my sincerest gratitude. I would also like to thank Thorsten Berger and Daniel Strüber for their significant contribution to the survey instrument.

References

[1] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, vol. 2, no. 3, p. 160, 2021.

[2] A. X. Zhang, M. Muller, and D. Wang, "How do data science workers collaborate? roles, workflows, and tools," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW1, pp. 1–23, 2020.

[3] N. Nahar, S. Zhou, G. Lewis, and C. Kästner, "Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process," in *Proceedings of the 44th International Conference on Software Engineering*, pp. 413–425, 2022.

[4] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," *Advances in neural information processing systems*, vol. 28, 2015.

[5] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pp. 291–300, IEEE, 2019.

[6] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young, "Machine learning: The high interest credit card of technical debt," in *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*, 2014.

[7] C. Hill, R. Bellamy, T. Erickson, and M. Burnett, "Trials and tribulations of developers of intelligent systems: A field study," in *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 162–170, IEEE, 2016.

[8] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel, "Understanding and visualizing data iteration in machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), p. 1–13, Association for Computing Machinery, 2020.

[9] S. Idowu, D. Strüber, and T. Berger, "Asset management in machine learning: State-of-research and state-of-practice," *ACM Comput. Surv.*, vol. 55, dec 2022.

[10] P. J. Guo, *Software Tools to Facilitate Research Programming*. PhD thesis, Stanford University, May 2012.

[11] M. Haakman, L. Cruz, H. Huijgens, and A. van Deursen, "Ai lifecycle models need to be revised: An exploratory study in fintech," *Empirical Softw. Engg.*, vol. 26, sep 2021.

[12] C. Shearer, "The crisp-dm model: The new blueprint for data mining," *Journal of Data Warehousing*, vol. 5, no. 4, 2000.

[13] G. Ericson, W. A. Rohm, J. Martens, K. Sharkey, C. Casey, B. Harvey, and N. Schonning, "Team data science process documentation," *Retrieved April*, vol. 11, p. 2019, 2017.

[14] S. Idowu, D. Strüber, and T. Berger, "Asset management in machine learning: A survey," in *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pp. 51–60, 2021.

[15] H. Liu, Y. Wang, W. Fan, X. Liu, Y. Li, S. Jain, Y. Liu, A. Jain, and J. Tang, "Trustworthy ai: A computational perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 14, nov 2022.

[16] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P. W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O'Keefe, M. Koren, T. Ryffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askell, R. Cammarota, A. Lohn, D. Krueger, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, A. Zilberman, Noa Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T. K. Gilbert, L. Dyer, S. Khan, Y. Bengio, and M. Anderljung, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," 2020.

[17] H. R, J. H, and S. M. JI, "Robustness and explainability of artificial intelligence," *EUR 30040 EN*, no. KJ-NA-30040-EN-N (online), 2020.

[18] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, 2018.

[19] R. Isdahl and O. E. Gundersen, "Out-of-the-box reproducibility: A survey of machine learning platforms," in *2019 15th International Conference on eScience (eScience)*, pp. 86–95, 2019.

[20] C. Liu, C. Gao, X. Xia, D. Lo, J. Grundy, and X. Yang, "On the reproducibility and replicability of deep learning in software engineering," *Transactions on Software Engineering and Methodology*, vol. 31, pp. 1–46, oct 2021.

[21] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, "Towards crisp-ml (q): a machine learning process model with quality assurance methodology," *Machine learning and knowledge extraction*, vol. 3, no. 2, pp. 392–413, 2021.

[22] R. Tatman, J. VanderPlas, and S. Dane, "A practical taxonomy of reproducibility for machine learning research," 2018.

[23] C. Drummond, "Replicability is not reproducibility: Nor is it good science," *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*, 01 2009.

[24] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M.-A. Sicilia, "Traceability for trustworthy ai: A review of models and tools," *Big Data and Cognitive Computing*, vol. 5, no. 2, 2021.

[25] J. A. Kroll, "Outlining traceability: A principle for operationalizing accountability in computing systems," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 758–771, Association for Computing Machinery, 2021.

[26] M. A. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S. A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar, "Accelerating the machine learning lifecycle with mlflow," *IEEE Data Eng. Bull.*, vol. 41, pp. 39–45, 2018.

[27] D. Baylor, E. Breck, H.-T. Cheng, N. Fiedel, C. Y. Foo, Z. Haque, S. Haykal, M. Ispir, V. Jain, L. Koc, *et al.*, "Tfx: A tensorflow-based production-scale machine learning platform," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1387–1395, 2017.

[28] E. T. Barr, C. Bird, P. C. Rigby, A. Hindle, D. M. German, and P. Devanbu, "Cohesive and isolated development with branches," in *Fundamental Approaches to Software Engineering: 15th International Conference, FASE 2012, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2012, Tallinn, Estonia, March 24-April 1, 2012. Proceedings 15*, pp. 316–331, Springer, 2012.

[29] E. Bisong and E. Bisong, "Kubeflow and kubeflow pipelines," *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pp. 671–685, 2019.

[30] T. Weißgerber and M. Granitzer, "Mapping platforms into a new open science model for machine learning," *it - Information Technology*, vol. 61, 04 2019.

[31] M. Schlegel and K.-U. Sattler, "Management of machine learning lifecycle artifacts: A survey," *ArXiv*, vol. abs/2210.11831, 2022.

[32] L. Quaranta, F. Calefato, and F. Lanubile, "A taxonomy of tools for reproducible machine learning experiments," *AIxIA 2021*, 2021.

[33] J. Linåker, S. Sulaman, R. Maiani de Mello, and M. Höst, *Guidelines for Conducting Surveys in Software Engineering*. Department of Computer Science, Lund University, 2015.

[34] C. Noy, "Sampling knowledge: The hermeneutics of snowball sampling in qualitative research," *International Journal of Social Research Methodology*, vol. 11, no. 4, pp. 327–344, 2008.

[35] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[36] O. E. Gundersen, "The reproducibility crisis is real," *AI Magazine*, vol. 41, no. 3, pp. 103–106, 2020.

[37] A. Searle, *Introducing Research and Data in Psychology: A Guide to Methods and Analysis*. Routledge modular psychology series: Perspectives and research, Routledge, 1999.

[38] E. R. Babbie, *The practice of social research / Earl Babbie, Chapman University.* Singapore: Cengage Learning Asia Pte Ltd, 14th edition. ed., 2016.

VIII. Appendix

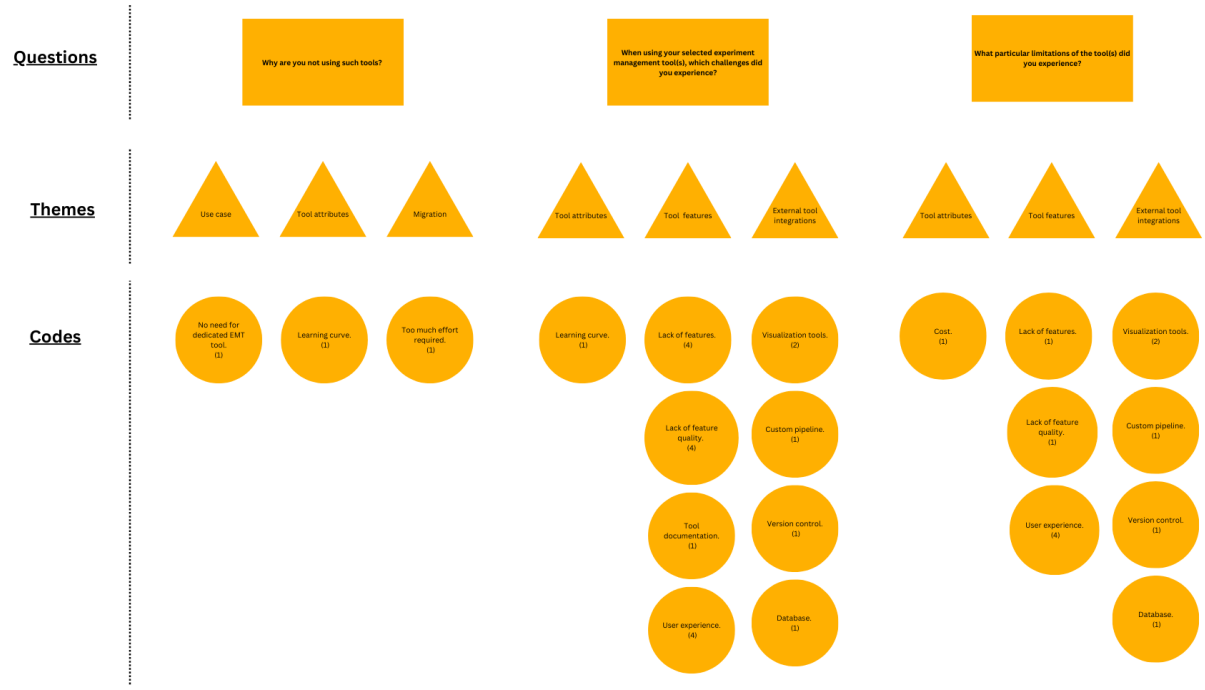

Fig. 13. Thematic overview of open-ended questions.